Research on Quantization Strategy Based on LGBM Algorithm

Group Name: Lihui

Group Member: Lihui Yan (ly2593)

April 22, 2023

Introduction

linear method： multicollinearity among factor

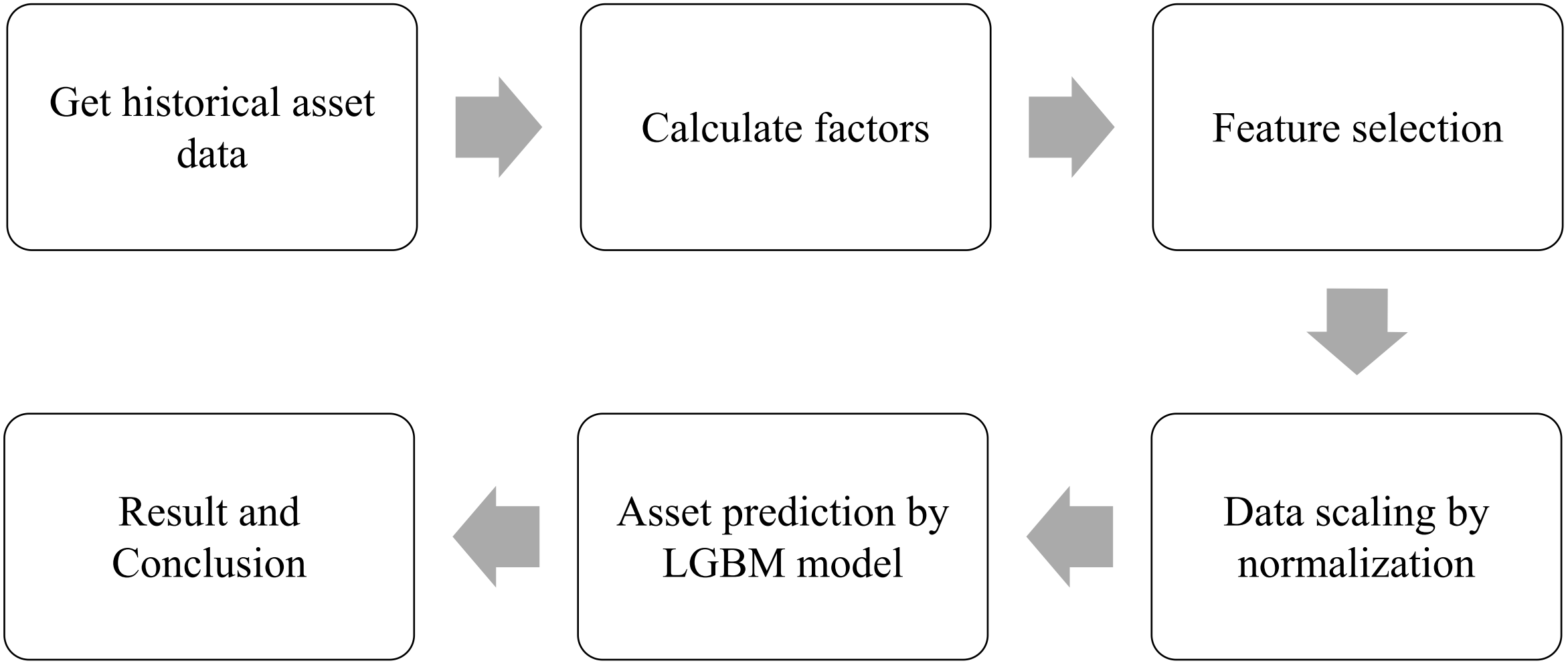
machine learning method:

1.deal with the multicollinearity problem

2.mine the nonlinear relationship between multiple factors.

When the same alpha factor is used, the machine learning algorithm may perform better.

Data and Methods



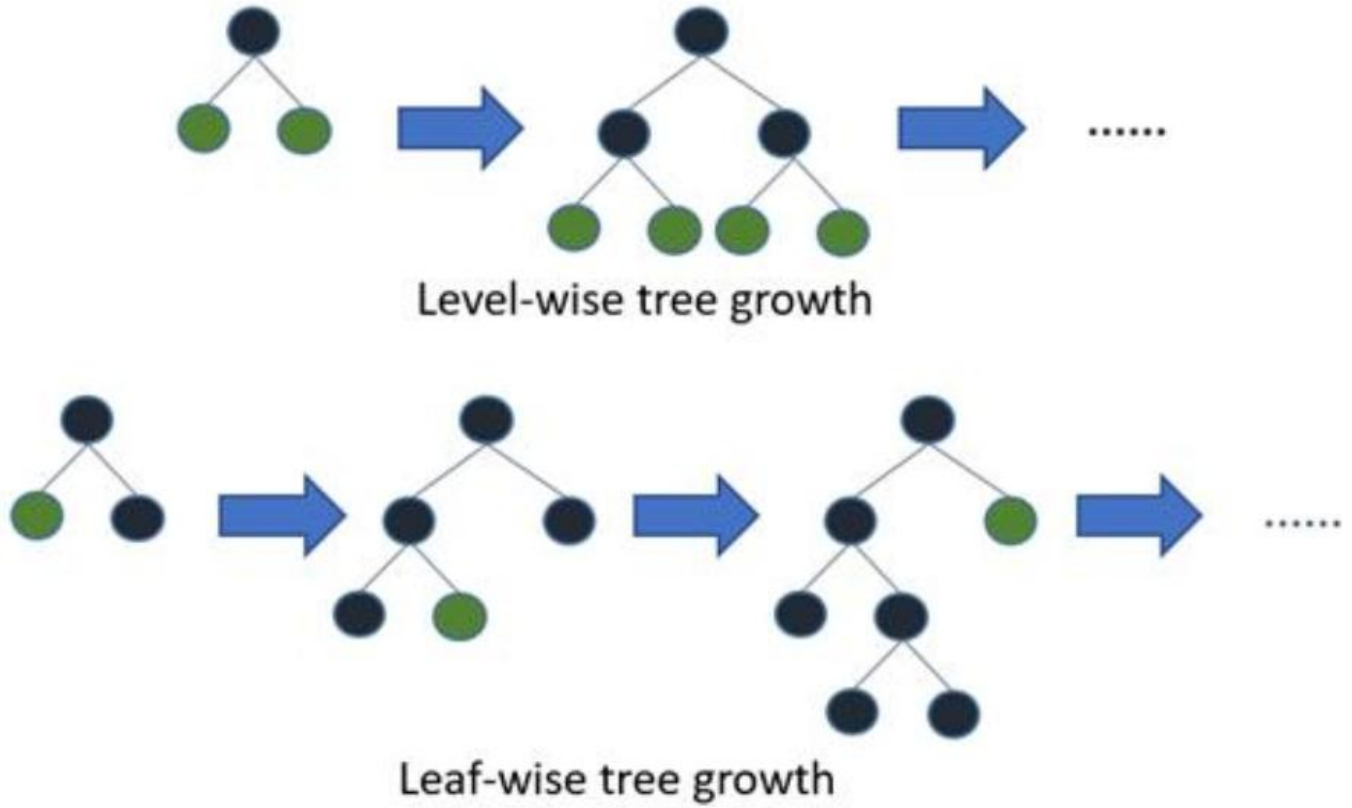
Data

The original data is scraped from yahoo finance. My forecast and investment decision include 50 stocks, 50 ETFs, and 10 crypto currencies. There are several vital indices that I keep track of or calculate each week.

1.Moving Average (MA)
2.Exponential Moving Average (EMA) $\text{EMA} = [\alpha \times \text{T Close}] + [1 - \alpha \times \text{YEMA}]$
3.Moving Average Convergence (MACD)
$\text{MACD} = [0.075 \times \text{EMA of Closeprices}] - [0.15 * \text{EMA of closeprices}]$
4.Total Volatility
5. Total Skewness
6. CCI
7. BIAS
8. KDJ
9. RSI

Model

Microsoft proposed the Light GBM model. It performs excellently in processing large-scale data.



Method

After calculating each factor, I need to preprocess all factors. Namely, remove the extremum, fill in the missing values, and finally standardize.

1.Extreme value processing(MAD5)

2. Missing value processing(mean)

3.Standardization(z-score)

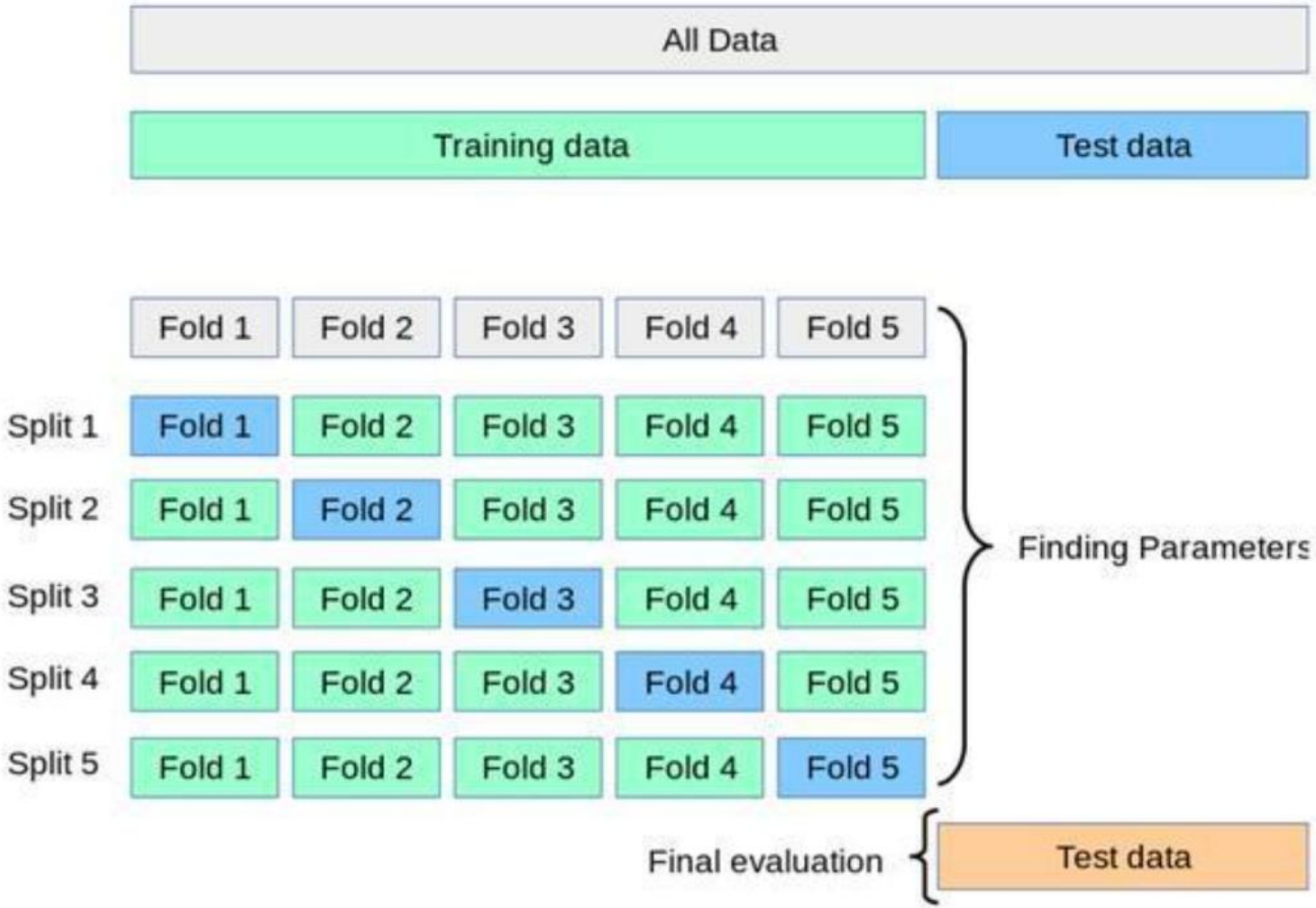
4.Label

Due to the use of a classification model, all assets are labeled. On the weekend of each week, select the top 20% of the next week's earnings as the sample with the highest rise (y=2), the assets ranked between 20% and 40% as the second most rising sample (y=1), the assets ranked between 40% and 60% as the third most rising sample (y=0), and the assets ranked between 60% and 80% as the fourth most rising sample (y=-1), assets ranked between 80% and 100% are used as the sample with the least increase (y=-2).

Method

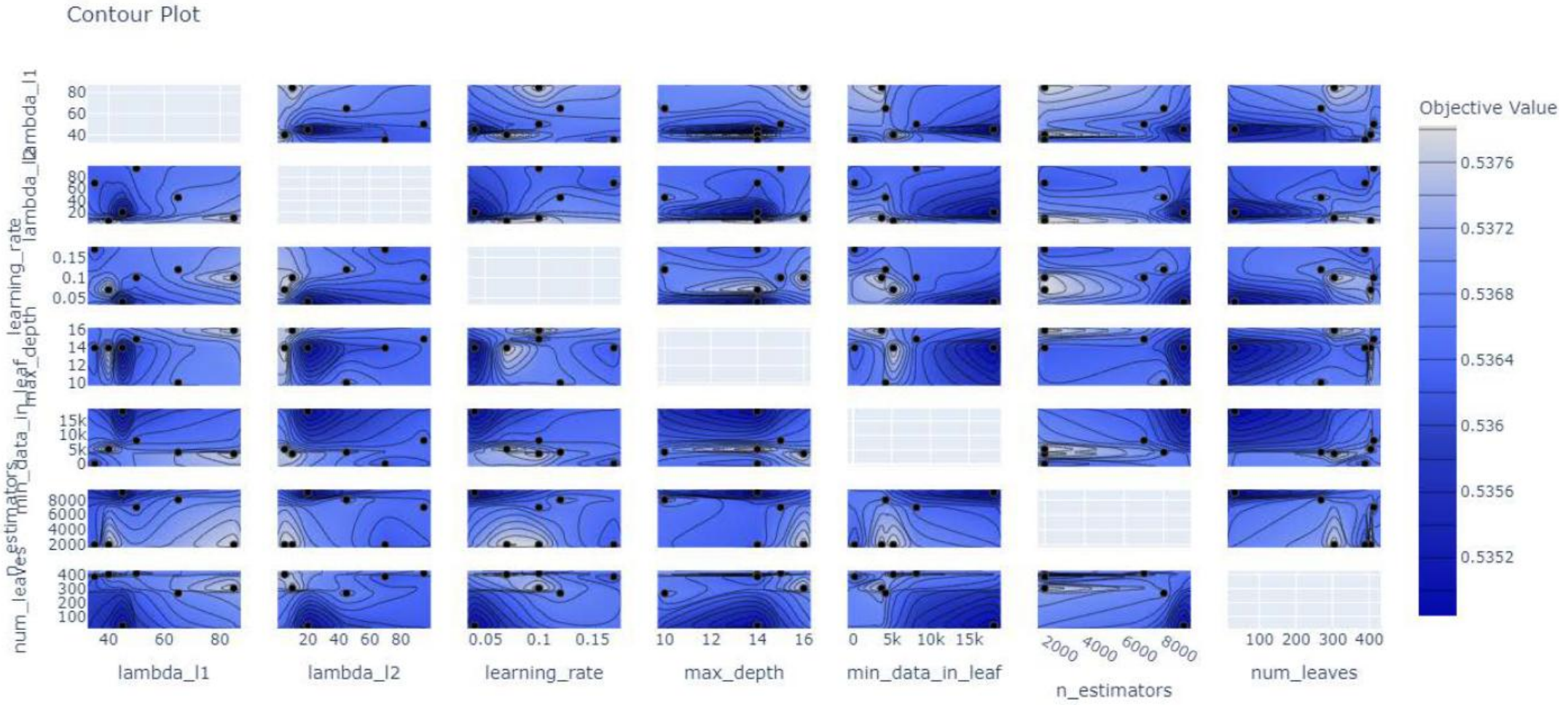
5.select the optimal parameters

Next, use K-fold cross validation to select the optimal parameters. Using a 50% interactive validation method, the best parameter of the average AUC of the validation set is selected as the optimal parameter of the model.



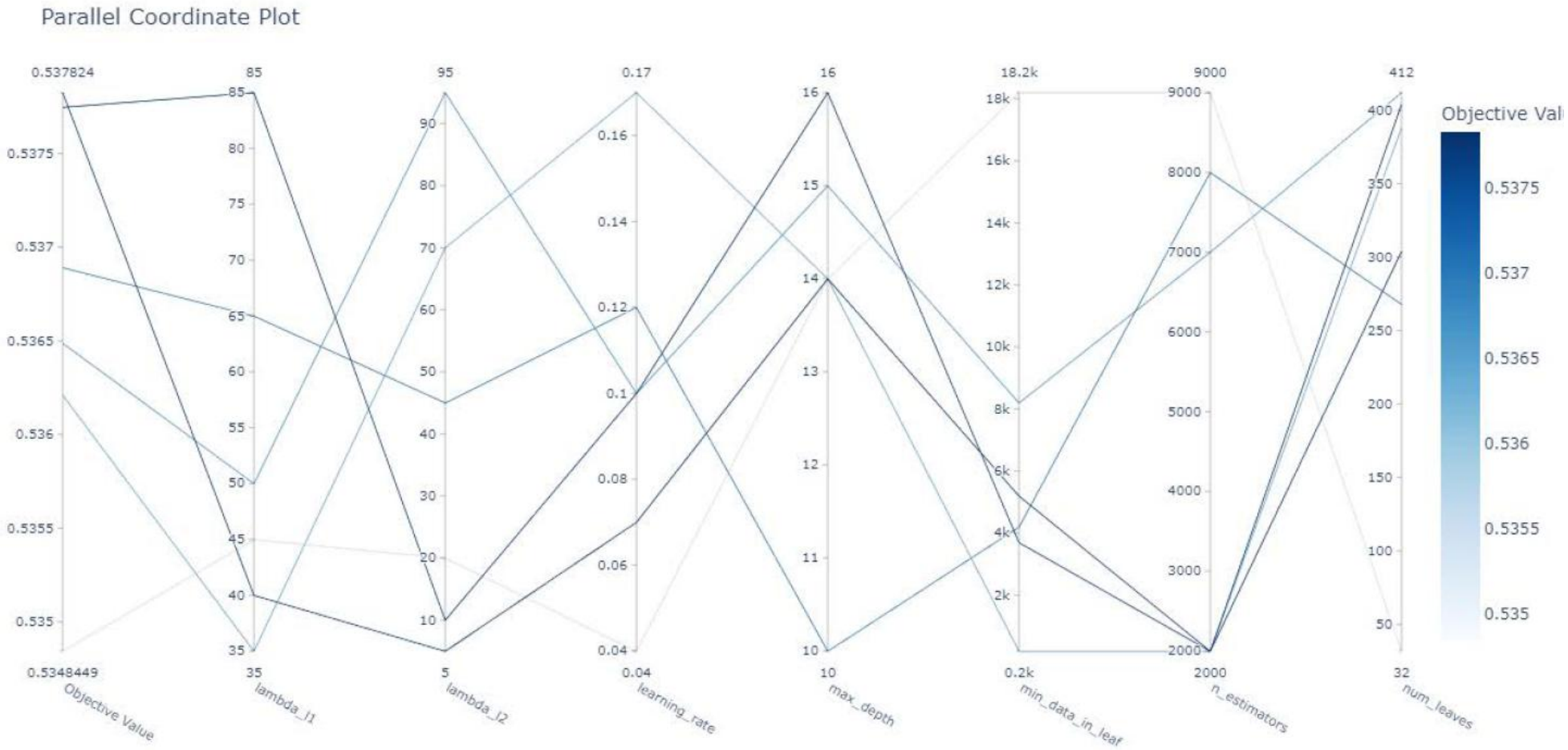
Method

The following figure shows the performance distribution of parameter optimization.



Method

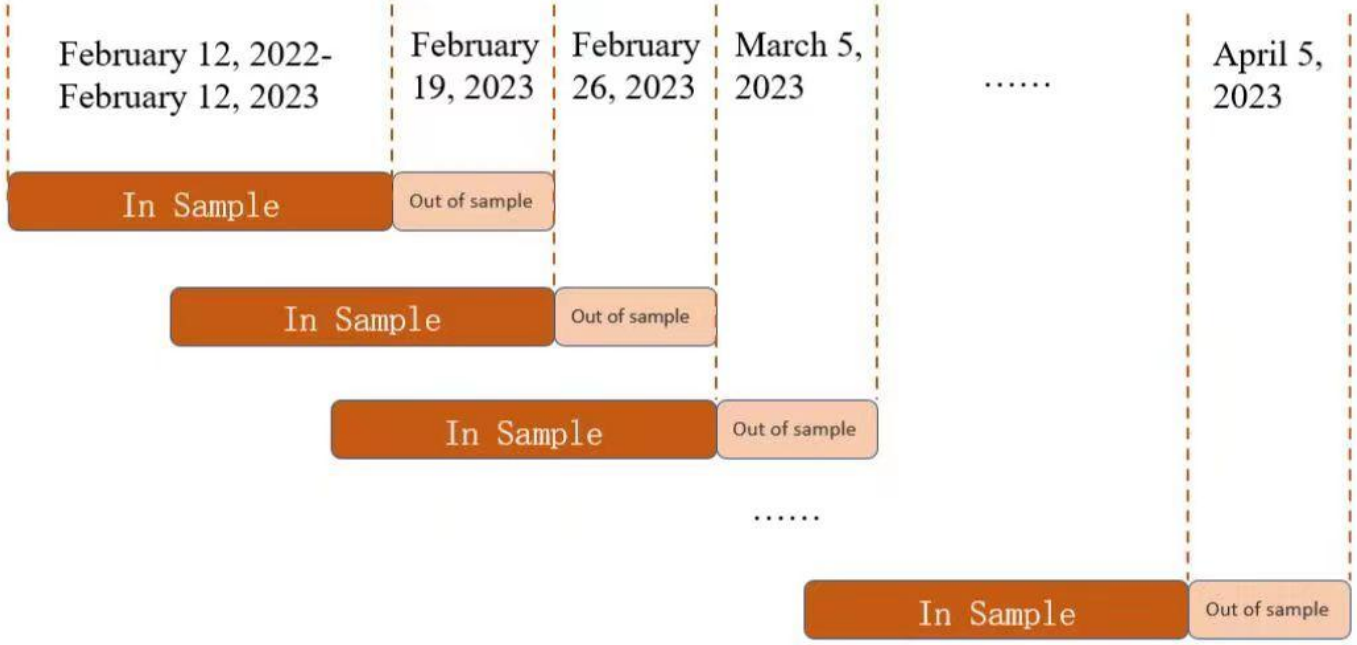
The following figure shows the parameters used to train the model every weekend.



Method

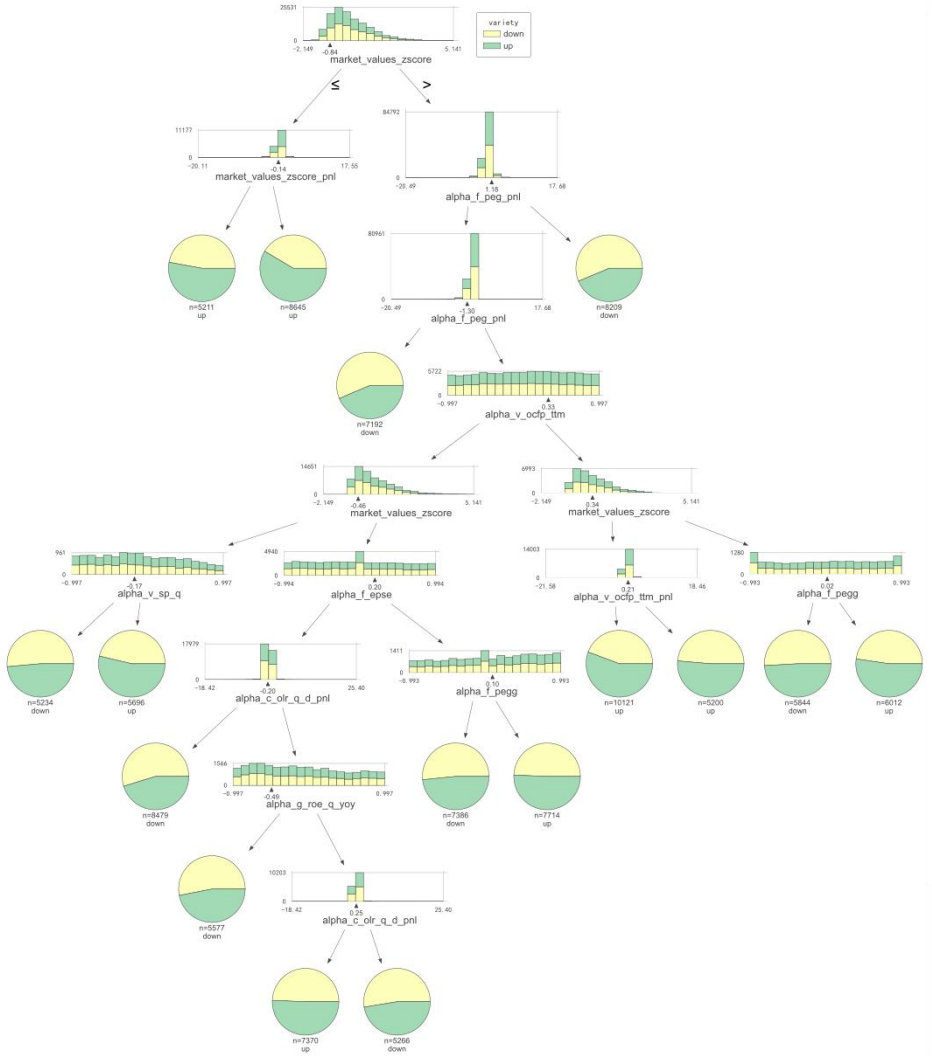
6. Training the model

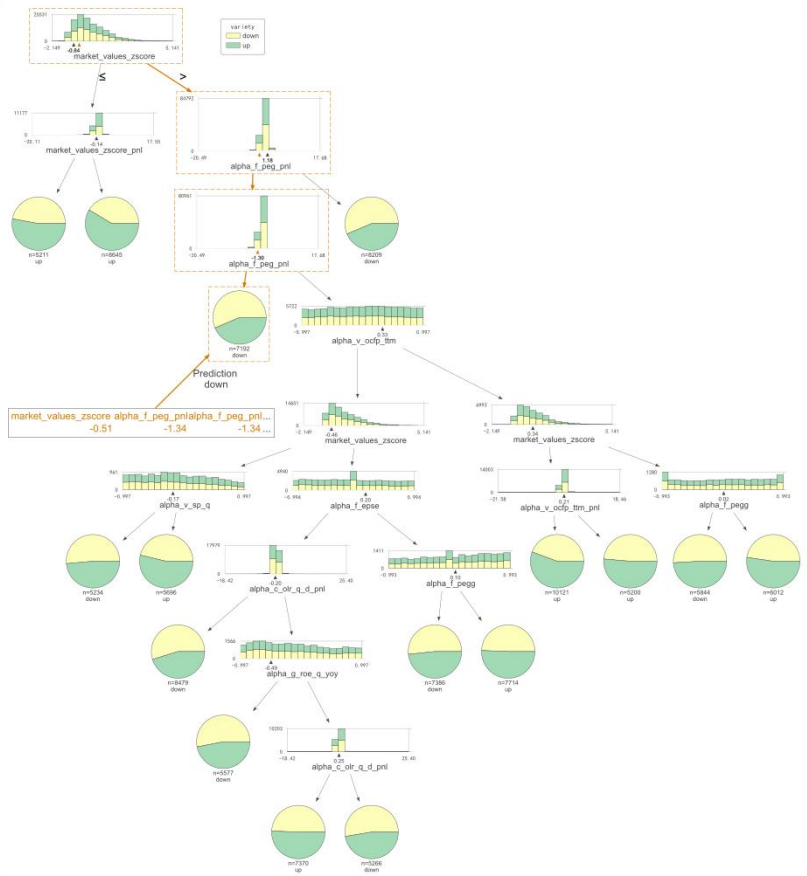
Then, using the method of rolling training the model, the samples with the current date pushed forward by past weeks are merged every weekend (each period is about one year of samples), and the optimal parameters mentioned above are used for rolling training.



Method

After determining the optimal parameters, train the machine learning model and input the feature values of all stocks into the trained model to calculate the probability of the model being in each classification for the next period of assets. I use the probability of the asset being predicted as Category 5 as the weight of the asset.





Results & Discussion

	forecasts_rank	decisions_rank	overall_rank
Feb 12	2	7	4.5
Feb 19	2	9	5.5
Geb 26	3	2	2.5
Mar 05	2	9	5.5
Mar 12	1	5	3
Mar 26	1	3	2
Apr 02	5	10	7.5
Apr 09	8	6	7

As for ranking performance, the strategy yielded satisfying results before Mar 26th. However, the strategy yielded unsatisfying results after Apr 2nd. This may because the model did not control the complexity during training, resulting in a overfitting problem.

Thanks for listening